공학석사학위논문

시공간 주의집중을 갖는 이중 흐름 행동인식 신경망

Two-stream Networks with Spatio-Temporal Attention for Action Recognition

2019년 8월

서울대학교 대학원

컴퓨터공학부

박 슬 기

공학석사학위논문

시공간 주의집중을 갖는 이중 흐름 행동인식 신경망

Two-stream Networks with Spatio-Temporal Attention for Action Recognition

2019년 8월

서울대학교 대학원

컴퓨터공학부

박 슬 기

요 약

오늘날 활발한 심층 신경망 연구와 데이터 저장 및 처리 기술 발달로 인해 이미지 뿐만 아니라 비디오와 같은 시간 흐름을 가진 대용량 데이터에서 다양한 인식 문제를 수행하는 연구가 더욱 더 많은 관심을 받고 있다. 그 중에서도 이중 흐름 신경망은 처음으로 신경망을 통한 학습이 기존의 수작업으로 뽑은 특징보다 (hand-crafted features) 좋은 성능을 보여준 이후로, 비디오 행동 인식에서 주류 아키텍쳐로 자리잡았다. 본 논문에서는 해당 아키텍쳐를 확장하여 비디오에서 동작 인식을 위해 독립적으로 훈련된 이중 흐름 신경망에 시공간 주의집중을 주는 아키텍쳐를 제안했다. 본 논문에서는 cross attention을 통해 기존의 독립적인 신경망에 상호 보완적인 학습으로 성능 향상을 유도했다. HMDB-51의 표준 비디오 행동인식 벤치 마크에서 본 논문의 아키텍쳐의 성능을 실험하였으며, 기존의 아키텍쳐보다 개선된 성능을 얻을 수 있었다.

주요어: 행동인식, 시공간 주의집중, 이중흐름신경망

학번: 2017-26622

차 례

표 차례

그림 차례

제 1 장 서론

비디오에서 인간의 행동을 인식하는 것은 도전적인 과제로, 컴퓨터 비전 연구자들의 많은 관심을 받고 있다 [1–8]. 고정되어 있는 이미지 분류에 비교하여 비디오 영상의 경우는 시간에 따라 등장인물의 동작이 달라진다. 이러한 다양한 동작이 중요한 단서를 제공할 수 있기 때문에 비디오 행동 인식에서 시간적 정보를 사용하는 것이 중요하다. 따라서 [1]에서는 공간적 요소 뿐만이 아닌 시간적 요소를 함께 고려하는 이중 흐름 합성곱 신경망 (Two-stream Convolutional Network) 을 제시하였다. 공간적 신경망(Spatial Network)에서는 비디오 프레임(이미지)를 처리하여 시각적 인식에 관여하며, 시간적 신경망(Temporal Network)에서는 프레임 사이의 움직임의 변화를 나타내는 옵티컬 플로우(Optical Flow)를 처리하여 움직임과 변화를 인식한다. 이중 흐름 합성곱 신경망 (Two-stream Convolutional Network)은 시각적 정보와 시간적 정보를 각각의 신경망에서 독립적으로 처리한 후, 마지막 단계에서 행동에 대한 점수를 합산한다. 이러한 이중 흐름 합성곱 신경망이 기존에 수작업으로 계산하여 뽑은 특징을 (hand-crafted features) 사용하는 방식보다 효과적임을 보인 이후로, 오늘날 가장 성공적인 비디오 행동 인식 프레임웍으로 자리잡게 되었다.

그런데, 공간적 신경망(Spatial Network)과 시간적 신경망(Temporal Network)은 각각 서로 다른 데이터를 처리하고 마지막에 그 점수를 합산할 뿐, 이전 학습 단계에서는 상호작용하지 않는다. 이러한 독립적인 이중 흐름 합성공 신경망은 UCF-101 데이터 [9]와 같이 '축구', '서핑', 등등 상대적으로 행동이 명확하게 구분될 때는 단순한 신경망 구성으로도 꽤 좋은 성능을 갖는다(88% [1]). 하지만 HMDB-51 데이터 [10]의 경우, 미세한 인간의 행동과 관련된 부분이 많아 상대적으로 어렵다 (59.4% [1]). 예를 들어 HMDB-51 데이터 중 '씹다(chew)'와 '미소짓다(smile)'을 구분할 때, 사람의 얼굴이 나온 고정 이미지로는 공간적 신경망(Spatial Network)에서 차이를 찾기 어렵다. 이 때, 시간적 신경망(Temporal Network)에서 얻은 정보를 통해 턱과 볼이 많이 움직이고 있으니 그 부분에 집중해야한다는 점을 알 수 있다면, 이

(a) RGB (b) Temporal Attn (c) Optical Flow (d) Spatial Attn

그림 1.1: **시공간 주의집중.** 이 그림은 시간적 신경망과 공간적 신경망에서 추출한 Attention map이 어떻게 상대 신경망에 도움을 줄 수 있는지를 보여준다.

행동은 '씹다(chew)'라는 것을 잘 맞출 수 있을 것이다. HMDB-51 데이터의 또 다른 예로 '공을 차다(kick ball)'와 '차다(kick)'의 경우를 생각해보자. 이전 프레임과 비교해서 움직임을 알려주는 옵티컬 플로우의 경우, 두 행위 모두 한 발을 움직이고자 하는 부분은 비슷할 것이다. 이 때, 시각적 정보에서 '축구공'을 인식하여 이곳에도 주의집중할 필요가 있음을 알려준다면, 공을 차는 행위인지 그냥 발을 차는 행위인지 차이를 더 잘 인식할 수 있을 것이다.

그림 1.1에서 이러한 attention 예시를 보여주고 있다. 그림 1.1 의 (a, b)는 각각 씹다('chew') 에 해당하는 비디오 프레임 이미지와 시간적 Attention 이다. (b)에서 시간적 Attention 은 사람이 껌을 씹을 때 가장 많이 움직이는 구겨지는 턱과 입 모양에 attend 하도록 잡아준다. 한 편, 그림 1.1 의 (c, d)는 각각 공을 차다('kick ball') 에 해당하는 옵티컬 플로우와 공간적 Attention 이다. 사람의 몸통 뿐만 아니라, 발 앞에 놓여진 공에 대한 Attention map을 전달함으로써 행동을 인식하는 데 결정적인 차이를 제공할 수 있는 지점을 알려줄 수 있다.

위와 같은 관찰을 통해 본 논문에서는 이중 흐름 합성곱 신경망 을 개선하여 두 흐름 신경망이 서로의 정보를 주고 받을 수 있는 시공간 주의집중 (spatio-temporal attention) 신경망을 제시하고자 한다. 그림 1.1에서 예시한 주의집중 맵(attention map)을 학습하여 상대 신경망에 주의집중해야할 영역 정보를 주는 구조를 설계하였

다. 시공간 주의집중 (spatio-temporal attention) 신경망을 통해 비디오 행동인식에서 가장 많이 사용되는 두 개의 벤치마크 데이터(UCF-101, HMDB-51)에서 기존보다 개선된 성능을 얻을 수 있었다.

제 2 장 관련 연구

2.1 행동 인식에서의 이중 흐름 신경망

Deep Convolutional Network 는 최근 비디오 행동 인식 연구 분야에서 좋은 성능을 나타냈다. 그 중에서도 데이터를 이중 흐름 신경망(Two-stream ConvNet [1])에서 제안된 두 개의 서로 다른 경로에서 데이터를 처리하는 아이디어는 현재 많은 연구에서 기본적으로 사용하고 있는 성공적인 방법 중 하나이다. 이중 흐름 신경망(Two-stream ConvNet [1])에서 공간적 신경망(Spatial Network)은 정지된 이미지로부터 객체 (‘무엇’)를 인지하며, 시간적 신경망(Temporal Network)은 여러 개로 쌓인 옵티컬 플로우(Optical Flow) 로부터 움직임(‘어떻게’)을 인식한다.

그러나 이러한 이중 흐름 신경망(Two-stream ConvNet)은 한 번에 하나의 이미지, 또는 일련의 옵티컬 플로우를 처리하여 비디오의 장기적인 컨텍스트를 반영 할 수 없다. 따라서 TSN [2] 논문에서는 시간적 맥락을 제시함으로써 이중 흐름 신경망(Two-stream ConvNet [1])의 성능을 향상 시켰다. TSN [2]은 비디오 클립을 동등한 기간의 K 세그먼트(segment)로 나누고 각 세그먼트에 대해 이미지와 옵티컬 플로우를 추출하여 이중 흐름 신경망을 훈련한다. 그리고 각 신경망의 합의 모듈(Consensus)이 세그먼트(segment) 간의 결과를 결합하여 최종적인 비디오 레벨 예측이 이루어진다.

이러한 기존의 이중 흐름(Two-stream) 접근법에서는 마지막 융합 레이어에서 각 네트워크에서 얻은 예측 점수를 결합하기 전까지는 상대 네트워크에 영향을 미치지 않는다. 그러나 개별적인 정지된 장면이나 ‘무언가’의 움직임은 비디오 내에서의 미묘하고 복잡한 인간의 행동을 인식하기에 충분하지 않다. 이 때, 이미지의 어느 부분이 많이 움직이는 지 알고, 따라서 해당 부분에 대한 주의가 필요한지 알 수 있다면 공간적 신경망(Spatial Network) 학습에 도움이 될 것이다. 마찬가지로, 시간적 신경망(Temporal Network)은 해당 움직임이 발생한 상황에 대한 문맥적 이해를 갖

고, 실제 액션과 관련된 부분에 더 많은 주의를 기울일 때 잘 학습될 것이다. 특히 카메라 이동, 장면 전환 등 전체 픽셀에 상당한 움직임이 발생하는 경우 더 효과적일 것이다. 본 논문에서는 이러한 관찰을 바탕으로 기본 Two-Stream ConvNet [1]을 기반으로 두 스트림간에 상호 작용을 도입하여 성능을 향상시키고자 한다.

2.2 행동인식에서의 주의 집중(Attention)

최근 비디오 기반 행동 인식 연구에 주의 집중 메커니즘(Attention Mechanism)을 포함하는 연구가 증가하고 있다. 이중 흐름 FCAN (Two-Stream FCAN) [4]은 Cross link 레이어를 통해 시간적 신경망(Temporal Network)에서 공간적 신경망(Spatial Network)으로 Attention을 추가하는 Flow-guided Convolutional Attention Network를 제안했다. Cross link 레이어는 시간적 특징(temporal features)으로부터 Attention map을 생성하고 이것을 공간적 신경망(Spatial Network)에 상호 대응하는 레이어의 Activation map에 element-wise로 곱해준다. Two-stream FCAN [4] 은 초기 레이어에서 Attention map을 생성하는 반면, 본 논문에서는 더 깊은 레이어의 Activation map을 선호한다. 깊을 레이어일수록 앞의 레이어를 통해 축적된 더 복잡한 정보를 가지고 있기 때문이다.

ST-ResNet [3]는 딥러닝 아키텍쳐 중 가장 성공적인 아키텍쳐 중 하나인 ResNet [11]에 영감을 받은 residual connection을 두 신경망 사이에 넣어 시공간 상호작용을 고안했다. 시공간 상호작용을 통해 서로 보완적인 정보를 전달하고 성능 향상을 도모했다는 점에서 본 논문의 모티베이션과 가장 유사하다. 하지만 본 논문에서는 학습된 attention map을 입력으로 전달하는 반면, ST-ResNet에서는 중간 합성곱 레이어에서 상대 레이어로 residual을 전달한다. 또한, 본 논문의 상호작용은 양방향으로 설계된 반면, ST-ResNet에서는 시간적 신경망에서 공간적 신경망 방향으로 residual을 주입한다는 점에 차이가 있다.

Attention map 추출을 위해 본 논문에서는 Grad-CAM [12] 기법을 활용한다. Grad-CAM [12]은 학습된 네트워크가 왜 특정한 예측 결과를 나타내는지 시각적

으로 이해하기 위해 제안된 기법이다. Grad-CAM [12] Gradient 기반의 위치 측정(localization)을 통해서 예측된 결과의 이유를 차별적으로 잘 설명하는 히트맵을 만들 수 있다는 점에 착안하였다. Grad-CAM [12]을 통해 한 이미지가 주어졌을 때 학습된 네트워크가 특정 class로 판단하기 위해 더 영향을 미치는 픽셀의 weight 를 구할 수 있기 때문에, 이를 신경망 훈련시 Attention map으로 활용하였다. 신경망 훈련 후 테스트 할때는 label을 모르므로 Grad-CAM을 구할 수 없기 때문에, label 없이도 Attention map을 생성하도록 별도의 합성곱 신경망을 훈련하였다.

제 3 장 시공간 주의집중을 갖는 이중 흐름 행동인식 신경망

본 논문에서는 공간적 신경망과(Spatial Network) 시간적 신경망(Temporal Network)간의 상호 작용을 가능하게하는 시공간 주의집중 모듈을 제안한다. 그림 3.1에서 보듯이 각 네트워크에서 공간적 및 시간적 의미를 독립적으로 학습하고, 학습된 정보는 상대 네트워크의 입력값으로 전달된다. 공간적 신경망(Spatial Network)에서는 RGB이미지에 temporal attention을 주기 때문에 시간적으로 중요한 부분에 더 집중하며 이미지에서의 객체를 잘 인식할 수 있다. 시간적 신경망(Temporal Network)에서는 옵티컬 플로우에 spatial attention을 주기 때문에 움직임을 인식할 때

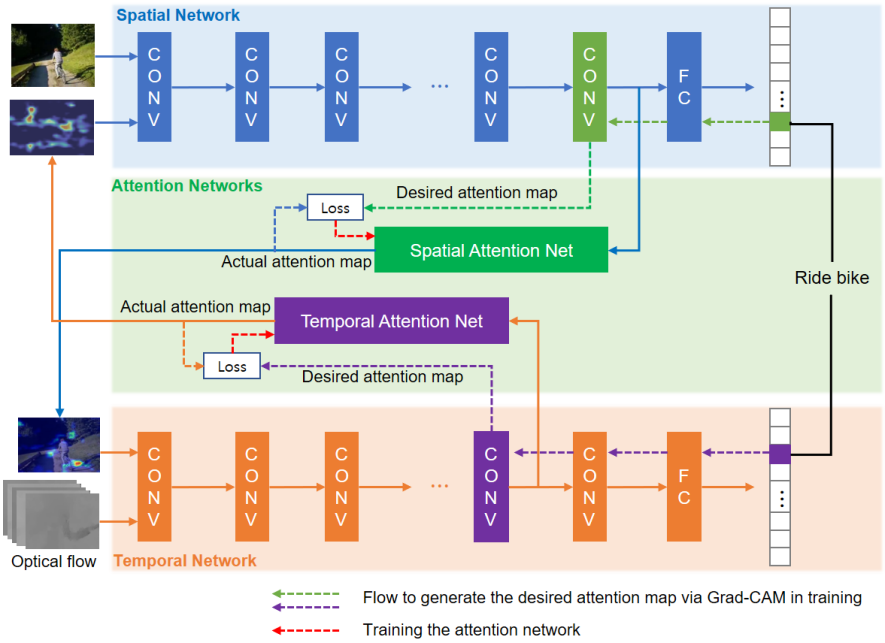


그림 3.1: **제안 아키텍쳐 개요.** 이 그림은 본 논문에서 제안한 아키텍쳐의 전체 흐름을 보여준다.

이미지 내에서 관련된 상황에 더 주의를 줄 수 있다. 이러한 교차 상호작용을 통해 각 신경망은 보다 복잡한 동작도 잘 인식할 수 있다. 학습 시에는 attention map는 label에 대한 gradient를 타겟 합성곱 레이어로 보내어 정확한 localization 유도할 수 있다. Attention map은 label을 이용하여 Grad-CAM으로 구할 수 있으나, test 시에는 label을 모르므로 label없이 Atention map을 추정할 수 있는 합성곱 신경망을 훈련할 수 있는 구조를 제안하였다.

뇌의 Ventral stream과 마찬가지로 공간적 신경망과(Spatial Network)은 RGB 이미지를 처리하여 객체('무엇')를 식별하도록 고안되었다. 시간적 신경망(뇌의 Dorsal stream)은 옵티컬 플로우(Optical Flow)를 처리하여 물체의 움직임 ('How')을 담당하도록 설계되었다. 옵티컬 플로우(Optical Flow)는 관찰자와 장면 사이의 상대적인 움직임에 의해 유발되는 물체, 표면 및 가장자리의 움직임의 패턴이다. 곧, 시간적 순서가 있는 두 이미지 사이의 움직임을 계산하는 방법으로 컴퓨터 비전 분야에서 널리 연구되고 있다. 비디오는 미세하게 변화하는 여러 이미지의 순서라고 볼 수 있는데, 이 프레임들을 바로 사용하는 것이 아닌, 옵티컬 플로우를 활용하여 움직임을 계산하는 것이 더 효과적임이 최근 연구에서 많이 밝혀졌다. 따라서 옵티컬 플로우를 처리하는 시간적 신경망에서는 학습이 진행됨에 따라 사람의 동작과 관련된 공간 및 시간적 특성이 학습된다.

이 때 각 네트워크는 학습된 정보를 attention map의 형태로 상대 네트워크에 입력으로 전달하므로 각각의 네트워크는 보다 복잡한 작업을 더 잘 인식 할 수 있게 된다. 다시 말해, 공간적 신경망(Spatial Network)이 학습된 attention map을 전달함으로써 시간적 신경망(Temporal Network)은 시각적으로 더 중요한 움직임에 주목하게 된다. 마찬가지로, 시간적 신경망(Temporal Network)은 물체가 어떻게 움직이는 지에 대한 정보를 공간적 신경망(Spatial Network)에 전달해주기 때문에 공간적 신경망(Spatial Network)는 풍부한 데이터 중 시간적으로 의미있는 부분의 이미지에 집중할 수 있게 된다. 이러한 두 네트워크 간의 동적 상호 작용은 인간 행동 인식의 정확성을 향상시킨다.

3.1 효과적인 주의집중 추출

그렇다면 신경망으로부터 어떻게 중요한 피쳐(feature)를 추출하여 주의집중 맵(Attention map)으로 활용할 수 있을까? 이 질문에 대한 해결책을 본 section에 서술한다.

아이디어. 직관적으로, 좋은 Attention map이란 다음 세 조건을 만족해야한다. 1) 입력 이미지에서 도움이 되는 영역을 localization 할 수 있고, 2) 관심 행동 카테고리와 관련되며 다른 카테고리와는 차별화되어야 하고, 3) 학습된 피쳐를 효율적으로 압축할 수 있어야 한다. 위의 조건을 만족하는 방법을 찾기 위해 본 논문에서는 깊은 신경망에서 시각적인 설명을 제공하는 Grad-CAM [12]에 주목했다. Grad-CAM은 타겟 액션 예측에 긍정적인 영향을 주는 영역에 대한 히트맵을 유도할 수 있고, 타겟 액션이 달라지면 히트맵도 달라진다(조건 1 & 2 만족). 또한, Grad-CAM을 사용하면 원하는 합성곱 레이어를 선택할 수 있으므로 깊은 신경망 레이어에서 더 복잡하고 밀집된 사이즈의 피쳐를 뽑아낼 수 있다 (조건 3 충족). 또한, 추가적인 모듈이나 네트워크의 수정없이 쉽게 계산할 수 있다.

Grad-CAM을 이용한 Attention map 추출. 본 논문에서는 마지막 합성곱 레이어를 타겟으로 했던 Grad-CAM을 확장하여 여러개의 합성곱 레이어에서 Attention map을 뽑도록 제안했다. 이를 통해 우리는 타겟 행동 y에 대하여 공간적 주의집중 맵(spatial attention map) T과 시간적 주의집중 맵(temporal attention map) S를 구할 수 있다. 타겟 행동 y의 예측 점수를 z_y, 추출하고자 하는 M 개의 타겟 합성곱 레이어의 feature map을 $L^{(1)}, \cdots, L^{(M)}$라고 하자. 이 때 m번째 합성곱 레이어의 feature map $L^{(m)}$ 은 C_m 개 채널과 가로 W_m, 세로 H_m 크기를 가졌다고 하자. $L_c^{(m)}$ 이 $L^{(m)}$의 총 C_m개 채널 중 c번째 feature map을 지칭한다고 할 때, 우리는 해당 feature map의 중요도 가중치 $\alpha_c^{(m)}$ 를 아래와 같이 구할 수 있다.

$$\alpha_c^{(m)} = \frac{1}{W_m H_m} \sum_i^{W_m} \sum_j^{H_m} \frac{\partial z_y}{\partial L_{cij}^{(m)}}.$$

위에서 구한 중요도 가중치를 통해 m번째 합성곱 레이어의 행동 차별적인 (action-

discriminative) localization map $T^{(m)}_{\text{GradCAM}}$은 다음과 같이 계산할 수 있다.

$$T^{(m)}_{\text{GradCAM}} = ReLU(\sum_{c}^{C_m} \alpha^{(m)}_c L^{(m)}_c). \tag{3.1}$$

ReLU를 적용하는 것은 원하는 액션을 예측하는 데 긍정적인 영향을 주는 피쳐만 고려함으로써 관련없는 부분을 제외하게 되어 더 localization을 잘 할 수 있도록 유도한다. 위에서 구한 localization map의 크기는 합성곱 레이어의 피쳐 크기인 $W_m \times H_m$ 이기 때문에 이를 입력 RGB 이미지 혹은 옵티컬 플로우의 사이즈와 같은 $W \times H$ 로 만들어주기 위해서 이중선형 보간법(bilinear interpolation)을 사용한다. 이 후, 여러 개의 합성곱 레이어에서 구한 localization을 결합하고 attention map 값이 0에서 1 사이 값을 갖도록 normalize 해줄 수 있다. 같은 방식으로 공간적 attention map S 도 구할 수 있다.

Attention map 학습. Grad-CAM으로 구한 attention map은 타겟 label을 주어야 구할 수 있으므로 신경망 훈련시는 사용할 수 있으나 label을 모르는 test sample에 대해서는 정확한 attention map을 구할 수가 없다. 이를 해결하기 위해 Test 시

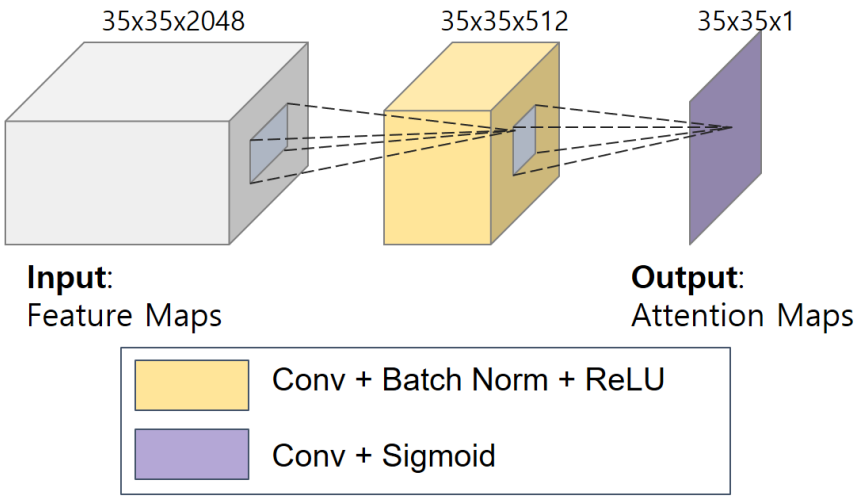


그림 3.2: **Attention Network 학습.** 이 그림은 attention map 학습을 위한 attention network의 세부적인 아키텍쳐를 보여준다.

사용될 attention map을 생성하는 합성곱 신경망(Attention Network)을 설계하여 훈련하였다. 그림 3.2에서 보듯이, 공간적/시간적 네트워크에서 추출된 feature map이 입력으로 들어가 같은 사이즈의 attention map을 출력으로 하는 신경망을 학습하게 된다. 전체 신경망은 두 개의 간단한 합성곱 레이어로 이루어져 있으며, 첫번째 합성곱 레이어에서 Batch normalization과 ReLU activation을 수행한다. 마지막 합성곱 레이어에서는 attention map 추출을 위해 합성곱 후 Sigmoid activation을 수행하여 attention map의 출력값이 0에서 1 사이의 히트맵을 가지도록 한다.

입력은 각각의 공간적 신경망 및 시간적 신경망의 합성곱 레이어에서 활성화된 피쳐맵($L^{(m)} \in \mathbb{R}^{W_m \times H_m \times C_m}$)을 사용하며, 출력은 attention map ($T, S \in \mathbb{R}^{W_m \times H_m}$)이 된다. 이후, Section 3.1 에서와 마찬가지로 이중선형 보간법(bilinear interpolation)을 사용하여 attention map의 사이즈를 $W \times H$로 만들어준다. 이 때의 합성곱 레이어는 Grad-CAM으로 attention map을 구할 때와 같은 합성곱 레이어($L^{(m)}$)를 사용한다. 학습 시 정답 출력으로 쓰일 desired attention map는 위에서 Grad-CAM으로 구한 attention map($T^{(m)}_{\text{GradCAM}} \in \mathbb{R}^{W \times H}$)을 사용하며, label을 알고 있는 training sample에 대해 8:2의 비율로 training set과 validation set으로 구분하여 학습시켰다. attention network를 통해 구한 attention map을 Test시 입력으로 줌으로써 공정한 Test를 할 수 있다.

Attention map 학습을 위한 목적 함수는 다음의 Binary Cross Entropy Loss로 나타낼 수 있다. attention network를 통해 출력된 시간적 attention map 을 $T(\in \mathbb{R}^{W \times H})$, desired attention map을 $T^{(m)}_{\text{GradCAM}}$라고 할 때,

$$\text{Loss}_{BCE} = \underset{w,b}{\text{Minimize}} \quad -T^{(m)}_{\text{GradCAM}} log T - (1 - T^{(m)}_{\text{GradCAM}}) log(1 - T).$$

이 때, w, b는 합성곱 신경망의 파라미터이다. attention map 학습 시 위와 같은 Binary Cross Entropy Loss를 사용하는 것이 좋은 이유는 다음과 같이 쉽게 증명 가능하다. 신경망은 목적 함수를 최소화하는 방향으로 학습이 진행되는데, 위의 목적함수가

최소가 되는 지점은 두 값이 같아질 때이기 때문이다.

$$
\begin{aligned}
&\frac{\partial \text{Loss}_{BCE}}{\partial T} = 0 \\
&\Rightarrow -T^{(m)}_{\text{GradCAM}} \frac{1}{T} - (1 - T^{(m)}_{\text{GradCAM}}) \frac{-1}{1-T} = 0 \\
&\Rightarrow -T^{(m)}_{\text{GradCAM}} (1-T) + (1 - T^{(m)}_{\text{GradCAM}}) T = 0 \\
&\Rightarrow \quad T = T^{(m)}_{\text{GradCAM}}.
\end{aligned}
$$

따라서 Binary Cross Entrop Loss 함수를 사용함으로써 두 attention map이 같아지도록 학습을 유도할 수 있다. 이 때 구현에서 학습을 위한 Optimization Algorithm으로는 Adam Optimization [13] 를 사용하였다.

3.2 행동패턴 학습과정

이 섹션에서는 행동패턴의 학습 과정을 기술하겠다. 이 학습과정은 그림 3.1에서 공간적신경망(Spatial Network)과 시간적신경망(Temporal Network)을 훈련시킴(가중치 갱신)으로 이루어 진다. 본 연구의 목적은 주어진 비디오 클립 V로부터 행동 $y \in \{1, \cdots, A\}$를 인식하는 것이다. 이를 위해 우선, 비디오 클립 V로부터 연속적인 RGB 이미지 $\{X_1, \cdots, X_n\}$ 와 옵티컬 플로우 $\{(u_1, v_1), \cdots, (u_{n-1}, v_{n-1})\}$를 추출한다. 이 때, 공간적 신경망(Spatial Network)는 RGB 이미지($X_i \in \mathbb{R}^{W \times H \times 3}$)와 시간적 attention map ($T_i \in \mathbb{R}^{W \times H}$)을 입력값으로 가지며, 이 때 이미지와 attention map은 모두 가로 W, 세로 H 크기를 가진다. 동시에 시간적 신경망(Temporal Network)는 m개 만큼 쌓아 붙인 옵티컬 플로우 $\{(u_i, v_i), \cdots, (u_{i+m-1}, v_{i+m-1})\} \in \mathbb{R}^{W \times H \times 2m}$ 와 공간적 attention map $S_i \in \mathbb{R}^{W \times H}$ 을 입력값으로 가진다. 이 때 두 이미지의 크기는 가로 W, 세로 H로, 위와 동일하다.

본 논문에서는 attention map을 임의적으로 0.1로 초기화하였다. attention map이 0에서 1 사이 값을 가지는데, 이 때 많은 값이 0 근처에 존재하기 때문이다. 휴리스틱하지만 이 방식이 이 방식이 1.0이나 0.5로 초기하는 것보다 안정적인 학습을

보였다. 이렇게 초기화된 시간적 attention map와 RGB image를 입력으로 하여 공간적 신경망(Spatial Network)를 학습시킨다. 한 차례의 학습이 끝난 후, Section 3.1에서 설명된 방법을 사용하여 학습된 공간적 신경망(Spatial Network)로부터 spatial attention을 계산한다. 이 후, 업데이트 된 spatial attention과 m개의 옵티컬 플로우를 통해 시간적 신경망(Temporal Network)를 학습시킨다. 학습이 끝나면 다시 temporal attention을 계산하고 기존의 것을 업데이트한다. 이러한 방식으로 신경망은 end-to-end로 학습되며, 전체적인 알고리즘은 Algorithm 1에 요약되어 있다.

비디오 행동 인식은 A개의 Class를 구분하는 문제이기 때문에 이중 흐름 신경망의 학습을 위한 목적 함수는 다음과 같이 정의할 수 있다. 각각의 공간적 신경망/시간적 신경망을 통해 예측된 행동을 p, 실제 label을 y라고 한다면, 각각 신경망의 목적 함수는 Categorical Classification 문제로, 다음과 같은 Cross Entropy Loss를 통해 학습 가능하다.

$$\text{Minimize} - \sum_{i=1}^{A} y_i log(p_i).$$

Algorithm 1: 이중 흐름 시공간 주의집중 행동인식 신경망 학습 과정

입력 : RGB 이미지 X_i,

m 개의 연속된 옵티컬 플로우 (u_i, v_i),

attention maps $S_i^{\text{GradCAM}}, T_i^{\text{GradCAM}}$

출력 : 행동 카테고리 y

초기화 Attention maps S_i, T_i 를 0.1 로 초기화

for *총 학습 횟수* **do**

- X_i 와 temporal attention T_i로부터 Spatial Net 학습.
- 학습된 Spatial Net으로 spatial attention S_i 업데이트 (GradCAM 방법 이용).
- $(\mathbf{u}, \mathbf{v})$ 와 spatial attention S_i로부터 Temporal Net 학습.
- 학습된 Temporal Net으로 temporal attention T_i 업데이트 (GradCAM 방법 이용).

end

두 네트워크의 결과를 합친다. 테스트시에는 Attention Net으로 추출한 attention map 이용.

제 4 장 실험

4.1 데이터셋과 구현 세부사항

데이터셋: 본 논문에서는 행동 인식의 벤치마크 데이터인 UCF-101 [9]과 HMDB-51 [10]에 대해 실험을 실시했다. UCF-101은 13,320개 비디오와 101개의 행동 클래스를 가지고 있으며, HMDB-51에는 6,766개의 비디오에 51개 행동 카테고리에 대한 주석이 달려있다. 두 데이터는 모두 3개의 training/testing split으로 나누어져 있다. 이 때, 비디오 데이터셋에서 옵티컬 플로우는 OpenCV에서 CUDA로 구현된 TVL1 알고리즘 [14]을 사용하여 추출한다.

학습 세부 과정: 공간적 신경망과 시간적 신경망의 구조는 모두 Inception-V3 [15]를 사용했다. 이 때, 학습 시간을 단축시키고 성능을 높이기 위해 두 신경망 모두 PyTorch [16]에서 이미지 분류를 위한 큰 데이터인 ImageNet [17]으로 미리 학습된 모델을 활용했다. 또한, 모서리 자르기(corner cropping), 다양한 크기로 자르기(multi-scale cropping), 임의 수평 반전(random horizontal flipping)과 같은 TSN [2]에서 소개된 학습 팁을 활용하였다. 이를 통해 학습 데이터에 변화를 부여함으로써 비디오 행동 인식의 데이터 부족 문제를 해결하고, 일반화 성능을 향상시킬 수 있었다.

공간적 신경망과 시간적 신경망은 Stochastic gradient descent 를 통해 학습시켰으며, 이 때 adaptive learning rate 알고리즘을 사용하여 특정 에폭 이후 learning rate을 10^{-1} 로 줄였다. HMDB-51 데이터의 경우, 초기 learning rate을 10^{-3}으로 학습 후 30 에폭, 250 에폭 이후에 각각 10^{-1}로 줄일 때 좋은 성능을 유도할 수 있었다. 공간적 신경망과 시간적 신경망의 학습 속도를 증진시키기 위해, 매번 attention map을 계산하기보다 0.1로 초기화한 attention map으로 일정 에폭 학습하여 기존의 이중 흐름 신경망과 유사한 성능을 얻은 후, Section 3.1의 방법을 사용하여 attention map

을 추출했다. 다시 말해, 두 단계로 나누어 Step 1에서는 초기화된 attention map과 RGB/옵티컬 플로우로 성능이 saturation될 때까지 학습하였다. 이후 이렇게 학습된 모델에서 attention map을 추출하고, Step 2에서 추출된 attention map을 넣어서 같은 모델을 다시 학습하였다. 그림 4.1에서 보는 것처럼 이러한 2단계 트레이닝이 end-to-end 방식으로 매번 attention map을 뽑는 것보다 더 빨리 트레이닝되며 attention map을 계산하는 시간을 줄일 수 있어 효율적이었다.

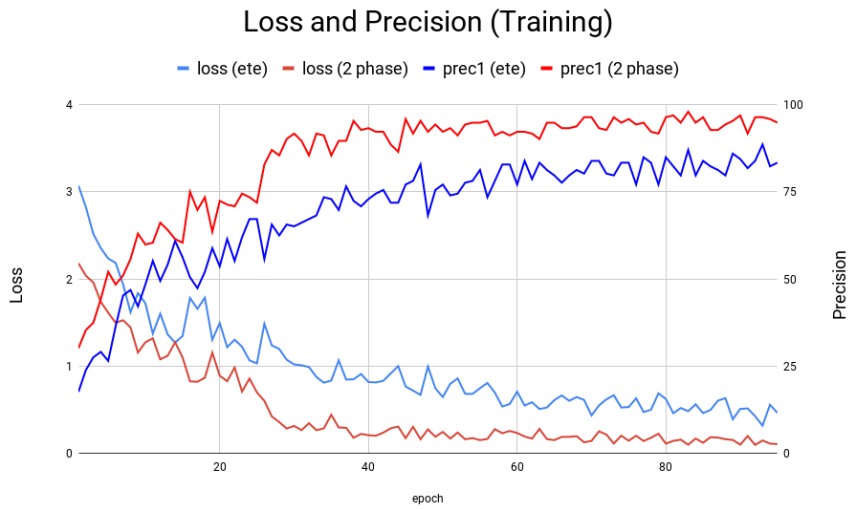


그림 4.1: **End-to-end 학습 비교.** 이 그림은 Training 단계에서 End-to-end Training과 2 phase training의 Loss/Precision 차이를 보여준다.

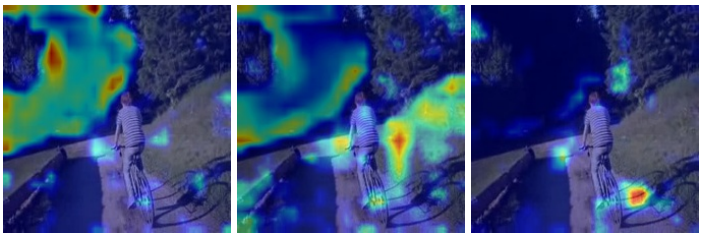
Grad-CAM [12] 방법은 한 합성곱 레이어를 대상으로 제시되었지만, 본 논문에서는 Inception-V3에서 Fully-connected 레이어 직전의, 여러 개의 합성곱 레이어 및 pooling 레이어로 이루어진 인셉션 모듈(inception module)에서 attention map을 만들었다. Inception-v3의 경우 여러 개의 합성곱 레이어가 하나의 인셉션 모듈을 형성하기 때문에 각각의 합성곱 레이어를 따로 보는 것보다 인셉션 모듈 단위로 배운 것을 이용하는 것이 직관적이라고 판단했다. 또한, 마지막 블록으로 갈수록 더 액션 카테고리와 관련되고, 구체적인 좁은 범위로 어텐션 맵을 얻을 수 있기 때문에 마지

막 블록인 Mixed_7C 레이어에서 얻은 피쳐맵으로 attention map을 만들었다 그림 4.2에서 인셉션모듈에 따른attention map변화를 보여준다.

Inception-V3 모델에서 Fully connected 레이어 직전 인셉션 모듈인 Mixed 7c를 기준으로 그 앞의 모듈은 각각 Mixed 7b, Mixed 7a 이다. 일반적으로 신경망에서 초기에는 선과 같이 일반적으로 공유될 수 있는 단순한 패턴을 인식하다가 레이어가 깊어질수록 더 액션과 관련있는 복잡한 패턴이 학습되는 것이 알려져있다. 그림 4.2 (a), (b)에서 보듯이, Mixed 7a, Mixed 7b 모두 자전거를 타다('ride bike')에 대해 배경에 더 집중하는 더 넓은 범위의 attention map이 추출되었음을 볼 수 있다. 한편, 그림 (c)의 Mixed 7c 모듈의 경우 더 자전거 바퀴와 길과 같은 카테고리와 직접적으로 관련된 구체적인 부분에 attention map을 형성하는 것을 볼 수 있다. 따라서 Mixed 7c 모듈에서 추출하는 것이 원하는 타겟 액션 카테고리와 관련되면서 필요한 부분에만 attention을 줄 수 있는 attention map을 얻을 수 있다.

이렇게 업데이트된 attention map으로 각각의 신경망에 cross attention을 준 후, 다시 공간적 신경망과 시간적 신경망을 학습하였다. 이 후, 각각의 신경망에서 예측된 결과를 가중 평균하여 최종적인 예측을 수행했다.

테스팅. 공정한 비교를 위해 베이스라인 이중 흐름 신경망 모델 [1]에서와 같은 테스트 방법을 수행했다. 한 비디오에 대해 25개의 이미지 프레임과 옵티컬 플로



(a) Mixed 7a (b) Mixed 7b (c) Mixed 7c

그림 4.2: **인셉션 모듈에 따른 attention map 변화.** 이 그림은 선택한 인셉션 모듈에 따라 추출된 attention map이 달라지는 양상을 보여준다.

우 스택을 뽑아 전체 예측 결과를 평균내는 방식을 사용했다. 이 때, 미리 학습된 attention network로 추출한 attention map을 사용하였다.

4.2 성능 비교

Attention 효과 우선, attention map을 사용하는 효과가 있는지 HMDB-51 데이터에 대해 확인해보았다. 각 공간적, 시간적 신경망에서 attention map의 효과를 확인해볼 수 있다(표 4.1). attention map을 넣지 않은 baseline 모델([1, 2])에 비해 attention map을 넣은 경우 각각의 신경망에서 조금씩 성능 향상이 있음을 확인할 수 있다. 특히, 이 결과는 공간적 신경망(Spatial Net)과 시간적 신경망(Temporal Net)을 fusion할 때 더욱 성능이 향상된 것을 통해 attention map이 최종적인 행동 인식 예측에 큰 기여하는 것을 확인할 수 있다.

표 4.1: **Attention 효과**

	HMDB-51 (split 1)		
	Spatial Net	Temporal Net	Fusion
Two-Stream [1]	40.5%	54.6%	59.4%
TSN [2]	55%	63%	68.5%
Spatio-Temporal Attention	55.03%	63.9%	71.2%

두 벤치마크 데이터 성능 비교 마지막으로 UCF-101, HMDB-51 두 벤치마크 데이터에 대한 성능을 비교하면 다음 표와 같다(표 4.2). 두 벤치마크 데이터에서 모두 기본 이중 흐름 신경망(Two-Stream ConvNet [1] 보다 성능 향상이 있음을 확인할 수 있다. UCF-101 데이터의 경우 Temporal Segment를 통해 여러 시간의 데이터를 함께 고려하는 TSN [2] 모델의 성능이 더 좋았다. 한편 HMDB-51 데이터에서는 본 논문에서 제시한 시공간 주의집중(spatio-temporal attention)을 추가한 모델의 성능이 더 좋았다. 이를 통해 더 인간의 미세한 행동과 관련 있어 세밀한 주의 집중이 필요한

표 4.2: **두 벤치마크 데이터 성능 비교**

	UCF-101	HMDB-51
Two-Stream [1]	88.0%	59.4%
TSN (2 modalities) [2]	94.0%	68.5%
Spatio-Temporal Attention	92.7%	71.2%

경우에 본 논문의 아키텍쳐가 도움이 되는 것을 확인할 수 있다.

제 5 장 결론

본 논문에서는 이중 흐름 신경망에 시공간 주의집중 맵(spatil-temporal attention map)을 추가하는 비디오 행동 인식 모델을 제안했다. 기존의 베이스라인 모델에 시공간 주의집중 맵을 추가하는 것으로 통해 성능 향상을 기대할 수 있었다. 특히 이 성능 향상은 HMDB-51과 같이 미세하게 유사한 행동이 많아 구별하기 어려운 행동 인식에서 더 좋은 성능을 보여주었다. 한편, 성능 향상이 대용량 Kinetics 데이터에서 pre-training하거나 이중 흐름 신경망을 3D Convolution등 다양한 방식으로 pooling 하는 등 오늘날 최신 비디오 행동 인식 연구 [5]에 비해서는 크지 않다는 문제가 있다. 하지만, 본 논문의 아키텍쳐는 해당 연구에 비해 계산할 파라미터 수 증가가 크지 않고, 대용량 Kinetics data pre-training 없이 ImageNet Pretraining 되어 있는 모델을 바로 이용함으로써도 성능 향상이 있었다는 장점이 있다.

본 논문의 추가적인 성능 향상을 위해 더 다양한 아키텍쳐 개선과 연구를 진행할 예정이다. Grad-CAM [12] 방식의 attention map을 뽑는 방식은 비디오 데이터 특성상 한 비디오 안에 수 십장에서 수 백장 사이의 연속된 프레임을 계산하는 데 시간이 오래 걸리며, 테스트 시 Label을 모르기 때문에 별도의 Attention Network 학습이 필요하다. 이를 개선하여 Grad-CAM의 지도(supervision) 없이 Training 단계에서부터 Attention Network를 바로 학습할 수 있고, 비디오 데이터 특성에 맞게 더욱 빠르게 attention map을 추출할 수 있는 방법을 연구할 계획이다. 또한, 더 심플하거나 복잡한 데이터 특성에 맞게 attention map 적용을 달리하는 방법을 통해 UCF-101에서도 성능을 높이는 연구를 진행할 계획이다.

참고 문헌

[1] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc., 2014.

[2] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 20–36, 2016.

[3] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3468–3476. Curran Associates, Inc., 2016.

[4] An Tran and Loong-Fah Cheong. Two-stream flow-guided convolutional attention networks for action recognition. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 3110–3119, 2017.

[5] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733, 2017.

[6] Jiagang Zhu, Zheng Zhu, and Wei Zou. End-to-end video-level representation learning for action recognition. In *24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018*, pages 645–650, 2018.

[7] Jue Wang, Anoop Cherian, Fatih Porikli, and Stephen Gould. Video representation learning using discriminative pooling. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1149–1158, 2018.

[8] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7024–7033, 2018.

[9] Khurram Soomro, Amir Roshan Zamir, Mubarak Shah, Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, page 2012.

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep

networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[14] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-L^1 optical flow. In *Pattern Recognition, 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007, Proceedings*, pages 214–223, 2007.

[15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826, 2016.

[16] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

ABSTRACT

Two-stream architecture has been mainstream since the success of [1], but two important information is processed independently and not interacted until the late fusion. We investigate a different spatio-temporal attention architecture based on two separate recognition streams (spatial and temporal), which interact with each other by cross attention. The spatial stream performs action recognition from still video frames, whilst the temporal stream is trained to recognise action from motion in the form of dense optical flow. Both streams convey their learned knowledge to the other stream in the form of attention maps. Cross attentions allow us to exploit the availability of supplemental information and enhance learning of the streams. To demonstrate the benefits of our proposed cross-stream spatio-temporal attention architecture, it has been evaluated on two standard action recognition benchmarks where it boosts the previous performance.

keywords: Action recognition, Two stream network, Spatio-temporal attention

student number: 2017-26622